The interplay of sequence conservation and T cell immune recognition

Bresciani, Anne Gøther; Sette, Alessandro; Greenbaum, Jason; Nielsen, Morten; Arlehamn, Cecilia S Lindestam; Peters, Bjoern

Published in:
Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics

Link to article, DOI:
10.1145/2649387.2660843

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

The interplay of sequence conservation and T cell immune recognition

Anne Bresciani
Center for Biological Sequence Analysis
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
+45 4525 2425
annebresciani@gmail.com

Jason Greenbaum
La Jolla Institute of Allergy & Immunology
9420 Athena Circle
La Jolla, CA 92037
+1 (858) 752-6500
jgbaum@liai.org

Cecilia S. Lindestam Arlehamn
La Jolla Institute of Allergy & Immunology
9420 Athena Circle
La Jolla, CA 92037
+1 (858) 752-6500
cecilia@liai.org

Alessandro Sette
La Jolla Institute of Allergy & Immunology
9420 Athena Circle
La Jolla, CA 92037
+1 (858) 752-6500
sette@liai.org

Morten Nielsen
Center for Biological Sequence Analysis,
Technical University of Denmark,
2800 Kgs. Lyngby, Denmark
+45 4525 2425
mniel@cbs.dtu.dk

Bjoern Peters
La Jolla Institute of Allergy & Immunology
9420 Athena Circle
La Jolla, CA 92037
+1 (858) 752-6500
bpeters@liai.org

ABSTRACT
Predicting which peptides can elicit a T cell response (i.e. are immunogenic) is of great importance for many immunological studies. While it is clear that MHC binding is a necessary requirement for peptide immunogenicity, other variables exist that are incompletely understood.

In this study we examined the hypothesis that conservation of a peptide in bacteria that are part of the healthy human microbiome leads to a reduced level of immunogenicity due to tolerization of T cells to the commensal bacteria. This was done by comparing experimentally characterized T cell epitope recognition data from the Immune Epitope Database with their conservation in the human microbiome. Indeed, we did see a lower immunogenicity for conserved peptides conserved. While many aspects how this conservation comparison is done require further optimization, this is a first step towards a better understanding T cell recognition of peptides in bacterial pathogens is influenced by their conservation in commensal bacteria. If the further work proves that this approach is successful, the degree of overlap of a peptide with the human proteome or microbiome could be added to the arsenal of tools available to assess peptide immunogenicity.

Categories and Subject Descriptors
J.3 [Life and Medical Sciences]: biology and genetics, health.

BCB '14, September 20 - 23 2014, Newport Beach, CA, USA

http://dx.doi.org/10.1145/2649387.2660843

General Terms
Experimentation. Theory. Documentation.

Keywords
Immune epitope database, human microbiome, epitopes, sequence conservation, immunogenicity

1. INTRODUCTION / BACKGROUND
Several methods are available that can accurately predict binding of peptides to MHCI and MHCII molecules (Karosiene et al., 2012; Kim et al., 2012; Nielsen et al., 2008; Peters et al., 2006; Wang et al., 2008). Binding to an MHC molecule is an essential though not sufficient criterion for a peptide to be recognized by T cells as an immune target. Many aspects of why certain peptides are targeted by T cells why some do not remain unknown.

The immune Epitope Database (IEDB) (Vita et al., 2010) contains a catalog of experimentally characterized B and T cell epitopes with both positive and negative results. By mining these data, comparing which peptides create an immune response and which do not, we can learn more about the immune recognition patterns.

In this study, we examined the hypothesis that conservation of peptides from bacterial pathogens in the human proteome and/or the human microbiome impacts their ability to activate a T cell mediated immune response. Tolerance against peptides conserved in the human proteome is expected to result from negative selection against T cells reactive with self-peptides during T cell maturation while tolerance to peptides found in commensal microbacteria has been postulated to be the result of peripheral tolerance induced by regulatory T cells. However, the extent to which these mechanisms impact recognition of peptides from bacterial pathogens has not been systematically examined. Here we examined this hypothesis by correlating human immune recognition of epitopes available in the IEDB with sequence conservation data from the Human Microbiome Project (HMP). In

the process, we tackle the problem of how to quantify the degree of immunogenicity of a peptide given aggregate data from multiple studies. The result of this work in progress so far is that there is a significant impact of peptide conservation in the human microbiome on the immune reactivity of a peptide.

2. MATERIAL & METHODS

2.1 Data

2.1.1 Peptides tested for T cell immune recognition

The peptides used in this study have been queried from the IEDB (Vita et al., 2010). Peptides from bacteria for which responses were measured in humans were included in the study. We restricted the analysis to include only peptides tested for MHC class II restricted / CD4+ T cell responses. The exact query parameters were: Source Organism: Bacterium (id: 2, Synonym: bacteria), Immune Recognition Context: T Cell Response, Host Organism: Homo sapiens, MHC Class: positive assays were MHC-II restricted and negative assays were MHC-II restricted or restriction undefined. The obtained peptides were all restricted to be between 12 and 25 amino acids in length. This peptide set includes 24,490 unique peptides and contains information about source organism as well as how many times it was tested positive and/or negative.

2.1.2 The human proteome

To compare the IEDB peptides to the human proteome, a reference proteome dataset was downloaded from www.uniprot.org (Consortium, 2014) in fasta format including all proteins recognized as part of a reference proteome, including both canonical sequence and all isoforms, as well as both reviewed and unreviewed proteins.

2.1.3 The human microbiome

To compare the IEDB peptides to commensal bacteria, proteins encoded in annotated reference genomes from the Human Microbiome Project (HMP) (Peterson et al., 2009) were used. Protein sequences were retrieved in fasta format from the HMP Data Analysis and Coordination Center (http://www.hmpdacc.org/). Either the entire set of proteins from reference genomes were used or those limited to the 6 most highly represented body sites: airways, blood, gastrointestinal tract, oral, skin, and urogenital tract.

2.2 Methods

2.2.1 Immunogenicity score

The same peptide is often tested in multiple studies and individuals. It is not expected that even the most immunogenic peptide will be positive in every individual due to genetic difference (most of all in HLA composition) and differences in exposure history. In order to quantify immune reactivity, for each peptide a count was made of number of subjects tested, N, and number of subjects responded, R. For peptides submitted to the IEDB where no information exist on number of responded and tested donors the following values were assigned; R = 1, N = 1 if the peptide was noted as positive and R = 0, N = 1 if noted as negative.

A response frequency was then calculated as R/N. However, using this ratio a peptide found positive in 1/1 donors would be considered more immunogenic than a peptide responding in an unheard of 99/100 donors. In order to favor data from larger sample sizes, a correction of the response frequency was made. As we have a binomial distribution, the lower 95% bound of binomial confidence interval was chosen as the corrected response frequency.

The cumulative binomial distribution function is as following:

$$F(k;n,p) = \Pr(X \le k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \quad (1)$$

where k is number of successes, n is number of donors tested and p is probability of success. F is the probability of getting k or less success. We wanted to solve for p, when F = 0.95. F signifies a 95 % probability of getting at least k responded out of n tested, with p probability of success. If, for example, 5 responded out of 10 donors tested, the uncorrected response frequency (probability of success) is p = 0.5. When asking for what probability p would we observe at most 5 donors responding with 95% confidence (F = 0.95) we get p = 0.22 which is then the adjusted response frequency. If in contrast only 2 donors were tested and one was positive, the adjusted response frequency is 0.02.

As the function in equation (1) is a sum function, it is not easily solved analytically. Therefore different approximations have been made to make confidence interval for binomial distributions. Among others the Wilson score interval, which works for relatively small samples and/or extreme probabilities. The Wilson score interval is defined as:

$$\frac{1}{1+\frac{1}{n}z^2}\left(\pi + \frac{1}{2n}z^2 \pm z\sqrt{\frac{1}{n}\pi(1-\pi) + \frac{1}{4n^2}z^2}\right) \quad (2)$$

n being number of donors tested, π is the number of responded divided by number of tested (i.e. the uncorrected response frequency) and z being 1-½α of a standard normal distribution. In our case α = 0.95 because we want a 95 % confidence interval, and z is correspondingly 1.96. For peptides with higher numbers of responded and tested donors (N>50) the approximation above agrees with numeric solutions to equation (1) with less than 1% difference, but when we go to the extreme cases, where a normal distribution no longer can be assumed, the Wilson score interval overestimates the positivity. To avoid this issue a matrix with p values was made. To have a starting point the p value was first set to number of responded divided by number of tested donors (R/N) and adjusted until F = 0.95. This was done for all cases of $0 \le R \le 50$ and $1 \le N \le 50$. For instances with N > 50, equation (2) was used to calculate the adjusted response frequency. We use the corrected response frequency value as a measure for the peptide immunogenicity and refer to it as the peptide immunogenicity score.

2.2.2 Mapping epitopes

To compare the IEDB peptides to the human- and human microbiome proteomes, lists were made containing all unique peptides from the proteomes, to which the IEDB peptides were then mapped.

Oseroff et al. (2010) found that peptide isoforms of length 15 with 1 or 2 mismatches in most cases showed the same T cell reactivity in a given donor. Therefore, in our study peptides with up to 2 mismatches were allowed for 15mer peptides and considered to be 'mapping' to each other. This number was adjusted for the peptides of other lengths so that 2/15 mismatches were allowed

per amino acid in the peptide. No gaps were allowed. The mapping was done against 8 different lists; one list containing the human proteome, one containing the complete set of the human microbiome, and one for each of the six body sites chosen for closer analysis. The mapping was done using a custom script, which first generates all unique peptides in the epitope dataset and our lists, respectively, and then compare the two lists.

2.2.3 Statistical evaluation of results

For testing the significance of the difference in average positivity between body sites, a one-way ANOVA test was used.

For testing significance of the difference in average positivity between peptides mapping and not mapping to the human proteome and the complete human microbiome, student's t-test was used.

For the mapping to the complete human microbiome the peptides were also characterized as either positive or negative based on a cutoff in the immunogenicity score empirically found to be optimal at 0.04. For this a Fisher's exact test was used to test the significance of the distribution of positive and negative peptides mapping and not mapping.

3. RESULTS

3.1 Assembly of epitope dataset

The dataset assembled from the IEDB consists of a total of 24,490 peptides with unique sequences of which the main part originates from *Mycobacterium tuberculosis*. Table 1 shows the distribution of species in the dataset. Note that the total number of peptides in column 2 is higher than the number of peptides with unique sequences from the IEDB because 49 peptides are listed as originating from two different species in the IEDB and 2 peptides are listed for 4 species.

Table 1. The origin of the different peptides in the dataset. First column indicates the species, second column indicates how many peptides originate from the given species in the IEDB, and the third column shows the percentage of the total number of peptides.

Species from IEDB	**Number of peptides**	**% of peptides**
Mycobacterium tuberculosis	21039	85.63
Clostridium tetani	844	3.43
Brucella melitensis	500	2.03
Burkholderia pseudomallei	324	1.32
Mycobacterium leprae	287	1.17
Bacillus licheniformis	245	1.00
Leptospira interrogans	219	0.89
Bacillus anthracis	178	0.72
Mycobacterium bovis	149	0.61
Helicobacter pylori	102	0.42
Streptococcus pyogenes	90	0.37
Bacillus amyloliquefaciens	89	0.36
Bacillus lentus	86	0.35
Corynebacterium diphtheriae	54	0.22
Borrelia burgdorferi	44	0.18
Francisella tularensis	37	0.15
Streptococcus mutans	33	0.13
Bordetella pertussis	33	0.13
Pseudomonas aeruginosa	31	0.13
Neisseria meningitidis	26	0.11
Staphylococcus aureus	18	0.07
Escherichia coli	17	0.07
Haemophilus influenzae	16	0.07
Chlamydia trachomatis	15	0.06
Listeria monocytogenes	13	0.05
Others	82	0.33
Total:	24,571	100.0

Table 2. Breakdown of peptides mapping to proteomes. The table gives an overview of the total number of peptides in the dataset, how many of these peptides map to different lists, and how large the different lists are.

	Number of peptides	Number of proteomes in list
Total records from IEDB	36,619	-
Unique peptides from IEDB	24,490	-
Mapping to the human proteome	45	1
Mapping to the complete human microbiome	9,129	1,097
Mapping to airways	1,888	50
Mapping to blood	289	42
Mapping to gastrointestinal tract	1,454	354
Mapping to oral	1,959	193
Mapping to skin	3,036	114
Mapping to urogenital tract	8,192*	128

** The high number of peptides mapping to the urogenital tract, compared to the other body sites, are due to one Mycobacterium present in that list and the abundance of Mycobacterium tuberculosis peptides in our epitope dataset.*

3.2 Mapping to the human proteome

Given our mapping criteria, only 45 bacterial peptides with immune recognition information mapped to the human proteome. Figure 1 shows the distribution of the peptides. When calculating their immunogenicity score, the mean of peptides that mapped to the human proteome was 0.0233 and the mean of peptides not mapping was 0.0082. While we had expected a lower immunogenicity of peptides mapping to the human proteome, given the low number of peptides mapping, this dataset is simply not powerful enough to support or refute this hypothesis.

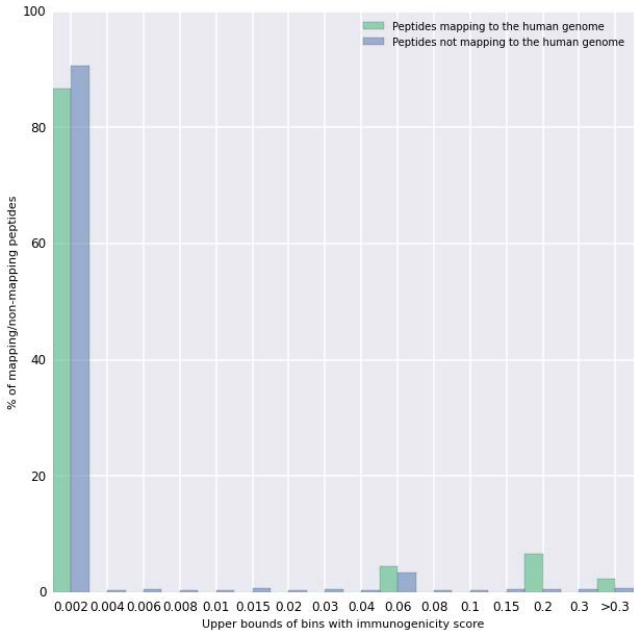

Figure 1. Comparing the distribution of immunogenicity scores for peptides mapping to the human proteome vs. those that do not. The plot shows the distribution of the peptides based on their immunogenicity score (positivity fractions) for the mapping to the human proteome. The peptides were divided into bins based on their positivity (immunogenicity score). The fraction of peptides mapping to the human proteome in each bin are indicated in green while those that did not map to the human proteome are indicated in blue.

3.3 Mapping to the human microbiome

Next, we mapped the bacterial peptides with immune recognition information to the human microbiome and found a much larger number of mapping peptides, namely 9,129. The distribution of the immunogenicity scores is shown in Figure 2. The average immunogenicity score of peptides that mapped to the human microbiome was 0.0073, while the peptides that did not map had a higher mean score of 0.0093. This difference is statistically significant (p-value = 0.0022, two-sided t-test). This suggests that bacterial peptides that are found in commensal bacteria present in the human microbiome have an overall lower ability to trigger an immune response.

3.4 Differential analysis of microbiota in separate body sites

Finally, we considered peptides mapping to microbial organisms found in different body sites separately. We considered the 6 body sites for which the largest number of reference genomes was available in the HMP project. Figure 3 depicts the number of peptides mapping. The p-value from the one-way ANOVA test, testing whether the average immunogenicity score for peptides mapping to each of the six body sites are different, is $4.182*10^{-14}$ which is highly significant on a 95 % level of significance. In the post hoc analysis examining which body sites contribute most to this difference, the standard error was calculated to be 0.0034, this was plotted in figure 3 and thus it can be seen that peptides mapping to microbiota found in the skin and urogenital tract have a significantly lower immunogenicity from those mapping to airways, blood, and oral, and also that peptides mapping to microbiota in the gastrointestinal tract have a significantly lower immunogenicity from blood and oral.

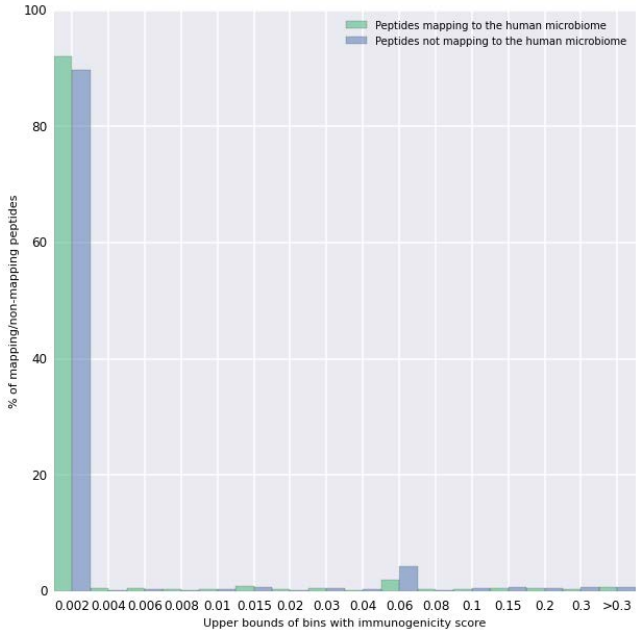

Figure 2. Comparing the distribution of immunogenicity scores for peptides mapping to the human microbiome vs. those that do not. As in Figure 2, the distribution of peptides mapping are shown in green while those that do not map are shown in blue.

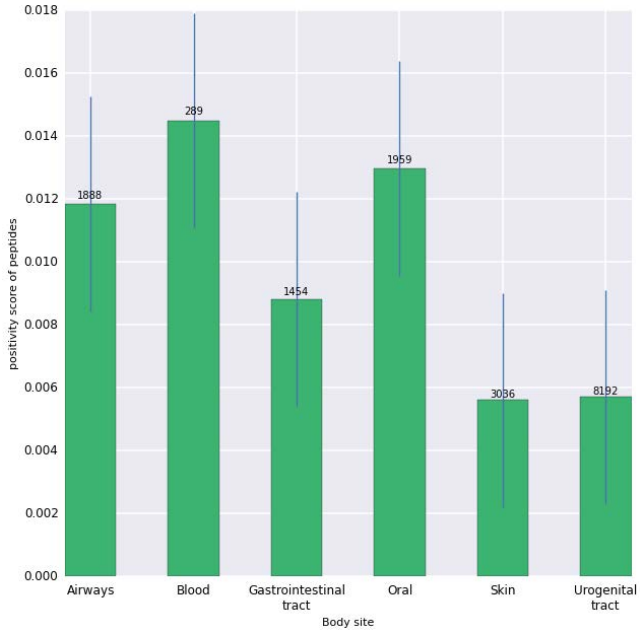

Figure 3. Immunogenicity of peptides mapping to microbiomes found in different body sites. The plot shows the differences in average immunogenicity score between the 6 different body sites. The number on top of each bar is the number of unique peptides mapping to that body site and the average positivity is calculated based on these. The error bars show the standard error of the differences between the groups.

4. DISCUSSION / CONCLUSION

The results of our work in progress are promising. Our starting hypothesis was that conservation of a peptide in the healthy human microbiome would lead to a reduced level of immune recognition due to tolerization to the commensal bacteria. This is supported by our results. We do see a lower average positivity for the mapping peptides than for the non-mapping peptides. This supported our theory and prompted us to investigate further.

Therefore we looked into the different body sites. It is known that the commensal bacteria in the gastrointestinal tract interacts closely with the immune system (Kamada et al., 2013) and so we would expect it to be different from that of other sites. Our analysis shows that it is significantly lower than blood and oral, but higher (although not significantly) than skin and urogenital tract. The skin acts a physical barrier and therefore one would not expect a close interaction with the immune system and hence not a high tolerization.

There are several caveats to our current analysis. Most importantly, the epitope dataset used in this study is largely dominated by the peptides originating from *Mycobacterium tuberculosis* (TB), which constitutes 85.63 % of the peptides (See table 1). This is mainly due to a whole genome study of the epitopes in Mycobacterium tuberculosis done by (Lindestam Arlehamn et al., 2013).

Having tested the entire genome for T cell reaction creates a dataset that is very comprehensive but hard to compare other organisms such as *Clostridium tetani*, where the peptides exclusively originates from the tetanus toxin, none of which are mapping to neither the human proteome nor the human microbiome. Since the entire TB genome were studied the peptides contain a larger proportion of negative peptides compared to other studies where only peptides assumed to have a T cell response were tested. It could influence the results that the average positivity of the TB peptides is lower than the rest of the dataset. Therefore the next step in this study is to analyze the TB data and the rest of dataset separately.

The results for the mapping of the human proteome are non-significant and furthermore the results actually show the opposite of what we would expect. It is known that there is a negative selection of self-antigens and so we expected to see that reflected in the results. However out of 24,490 peptides there are only 45 mapping peptides, which is too low to make strong conclusions. One approach we will test going forward is the use of multiple human proteomes to represent the inter-individual variability of what is considered a self-antigen, which might pick up additional mapping peptides.

Overall, we believe our analysis is a first step towards a better understanding how conservation of peptides in commensal bacteria shapes their recognition when present in bacterial pathogens. Given the inherent limitations when using public data generated for a completely different purpose than what we are examining in our study, our plan to address the caveats above is to formulate a testable hypothesis, and design a dedicated experiment to test it. If successful, the degree of overlap of a peptide with the human proteome or microbiome could be added to the arsenal of tools available to assess peptide immunogenicity, which have practical applications in peptide vaccine design and de-immunizing of protein therapeutics.

5. ACKNOWLEDGMENTS

The La Jolla Institute of Allergy and Immunology is supported by the National Institutes of Health National Institute of Allergy and Infectious Diseases, Allergy Contract no. HHSN272201200010C under the Immune Epitope Database and Analysis Program. Morten Nielsen is a researcher at the Argentinean national research council (CONICET).

6. REFERENCES

Consortium, T.U. (2014). Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. *42*, D191–D198.

Kamada, N., Seo, S.-U., Chen, G.Y., and Núñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. Nat. Rev. Immunol. *13*, 321–335.

Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics *64*, 177–186.

Kim, Y., Ponomarenko, J., Zhu, Z., Tamang, D., Wang, P., Greenbaum, J., Lundegaard, C., Sette, A., Lund, O., Bourne, P.E., et al. (2012). Immune epitope database analysis resource. Nucleic Acids Res. gks438.

Lindestam Arlehamn, C.S., Gerasimova, A., Mele, F., Henderson, R., Swann, J., Greenbaum, J.A., Kim, Y., Sidney, J., James, E.A., Taplitz, R., et al. (2013). Memory T Cells in Latent Mycobacterium tuberculosis Infection Are Directed against Three Antigenic Islands and Largely Contained in a CXCR3+CCR6+ Th1 Subset. PLoS Pathog. *9*.

Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., Buus, S., and Lund, O. (2008). Quantitative Predictions of Peptide Binding to Any HLA-DR Molecule of Known Sequence: NetMHCIIpan. PLoS Comput. Biol. *4*.

Oseroff, C., Sidney, J., Kotturi, M.F., Kolla, R., Alam, R., Broide, D.H., Wasserman, S.I., Weiskopf, D., McKinney, D.M., Chung, J.L., et al. (2010). Molecular Determinants of T Cell Epitope Recognition to the Common Timothy Grass Allergen. J. Immunol. Baltim. Md 1950 *185*, 943–955.

Peters, B., Bui, H.-H., Frankild, S., Nielsen, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., et al. (2006). A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules. PLoS Comput Biol *2*, e65.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., et al. (2009). The NIH Human Microbiome Project. Genome Res. *19*, 2317–2323.

Vita, R., Zarebski, L., Greenbaum, J.A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. (2010). The Immune Epitope Database 2.0. Nucleic Acids Res. *38*, D854–D862.

Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A., and Peters, B. (2008). A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. PLoS Comput Biol *4*, e1000048.